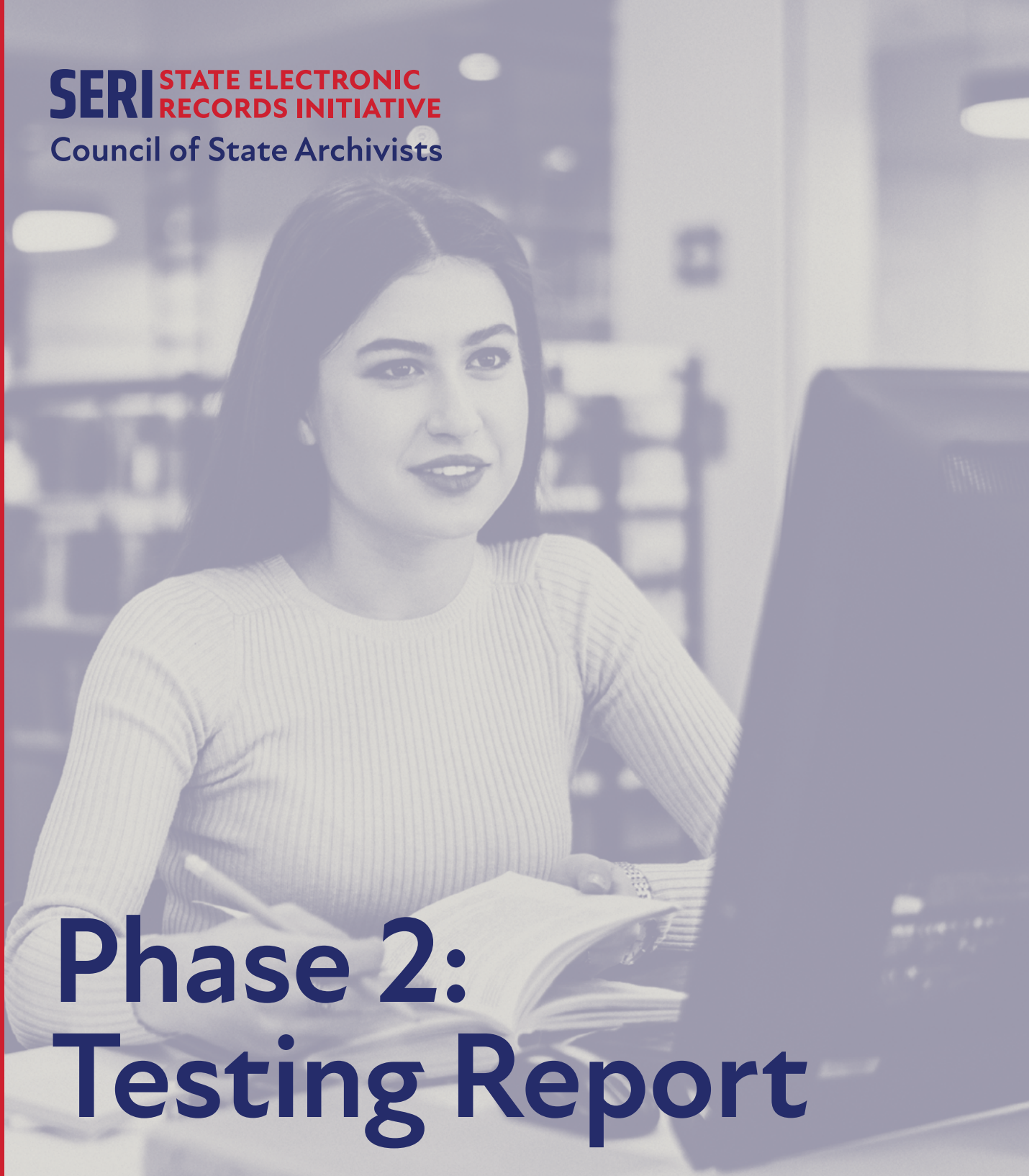
CoSA PREPARE: PREPARING ARCHIVES FOR RECORDS IN EMAIL

SERI STATE ELECTRONIC RECORDS INITIATIVE
Council of State Archivists



Phase 2: Testing Report

Prepared by Nick Connizzo

With editorial assistance from Michelle Gallinger, SERI Coordinator and the SERI Tools and Resources and Steering Committees

CoSA Council of State Archivists

SEP 2022

Contents

CoSA
Council of State Archivists

The Council of State Archivists (CoSA) is a nonprofit membership organization of the state and territorial government archives in the fifty states, five territories, and District of Columbia. Through collaborative research, education, and advocacy, CoSA provides leadership that strengthens and supports state and territorial archives in their work to preserve and provide access to government records. CoSA facilitates networking, information sharing, and project collaboration among its member organizations to help state and territorial government archives with their responsibilities for protecting the rights and historical documents of the American people.

EMAIL info@statearchivists.org
WEB www.statearchivists.org

Support for this publication was provided by the University of Illinois/Mellon Foundation for PREPARE: Preparing Archives for Records in Email, a program of capacity-building services for email management and preservation CoSA is providing to state and territorial archives as part of the State Electronic Records Initiative (SERI).

SERI STATE ELECTRONIC RECORDS INITIATIVE

About the State Electronic Records Initiative (SERI) The State Electronic Records Initiative (SERI) was established in 2011 to focus on improving management, preservation, and access to state and territorial government electronic records in all 56 states, territories, and the District of Columbia. SERI increases capacity and capability for state electronic records management and provides a robust community of practice for archival staff preserving electronic records. SERI is committed to supporting the communities of practice, documents, guidance, and other resources to help state and territorial archives achieve their missions and increase equitable access to state electronic records. SERI promotes knowledge of the digital lifecycle and requirements for the preservation of state electronic records. SERI advocates for increased awareness and preservation of electronic records. SERI is the flagship program of CoSA.

When empowered with the right tools, processes, and expertise, archivists should be able to more effectively communicate requirements for email preservation and incorporate them into existing email governance frameworks.

Background

The Council of State Archivists (CoSA) is providing capacity-building services for email management and preservation to state and territorial archives through the University of Illinois' Email Archives: Building Capacity and Community program, funded by the Andrew W. Mellon Foundation. In order to determine specific needs and interests, CoSA developed a needs assessment survey[1] and distributed it to its member organizations in May 2021.

The Needs Assessment Survey had two primary objectives: 1) to ascertain the extent of obstacles to effective email preservation broadly present within state and territorial government, and 2) to analyze existing capabilities to identify gaps in email preservation workflows for state and territorial archives. Obstacles to Email Preservation, as determined in the survey:

- **Lack of Policy Governance:** State and territorial archives lack written and fully-implemented policies for managing key digital lifecycle phases for email.
- **Lack of Standard Models for Preservation and Transfer:** Email records aren't being transferred to the archives as much as would be expected, likely caused by lagging procedural implementation. This interruption in the records lifecycle presents a growing concern.
- **Scale:** The volume of email is a scale issue that states are struggling to address. As above, lack of throughput in the records lifecycle is causing email datasets to continually grow in the best of circumstances, and be deleted for space and to reduce costs in the worst.
- **Technology:** State and territorial archives have had difficulty establishing effective and repeatable models due in some cases to a lack of necessary technology needed to manage email preservation. This includes small-scale needs (format converters, transfer applications) and large-scale needs (digital repositories).
- **Training/Experience:** Records producers and archives staff often are not adequately trained to manage email, let alone the its digital preservation.

Phase 2 of the PREPARE Project was conceived and designed to address some of these obstacles. Project staff, in consultation with the PREPARE Advisory Group, convened a cohort of digital archivists to test a sample email dataset under low-risk conditions with open-source tools. The tests provided insights on how successful processing models might look using these tools; additionally, archivists can provide recommendations to records producers and managers to ensure that valuable email makes it to the archives for long-term preservation.

Additionally, the Needs Assessment Survey Analysis and Report identified three areas as achievable within the scope of the PREPARE Project:

- **Preservation Pathways for Email Records:** State and territorial archives at lower levels of digital preservation capability lack defined procedural pathways to transfer permanent records from the custody of records producers to digital repositories.
- **Identification of Email Preservation Lifecycle Elements:** State and territorial archives require frameworks that address classification and taxonomy; archival appraisal and processing; transfer; storage and preservation; search and retrieval; and migration of email records. Traditional frameworks for records management and archival preservation can be outdated and require updates to address the scale and scope of email records.
- **Email Policy Adoption:** The assessment concluded that the application of email policy at the archives was helpful, but insufficient for optimal preservation. Controlling the creation, management, and description of email is necessary at the point of records-creation. Awareness of available tools and processing/preservation workflows will allow archivists to make informed decisions about what policies and procedures for managing

1 Council of State Archivists. CoSA PREPARE: Needs Assessment Survey Analysis & Report. 2021.

email are best suited to each jurisdiction's unique requirements.

To address these specific areas, the testing paradigm was designed with simple workflows that could be applicable to a wide variety of email preservation scenarios. Furthermore, volunteers were asked to specifically gather information about the software's efficacy related to the aforementioned areas in order to best judge its candidacy for inclusion within an email preservation framework.

Testing Method

Phase 2 of the PREPARE Project focused on the following outcomes:

- Provide a low-risk environment for digital archivists to test software for its inclusion in state and territorial archives' email preservation workflows. Low-risk, in this case, meant minimal impact on existing operations, records, or infrastructure.
- Offer an opportunity for digital archivists and records professionals to test available and affordable software tools and potential workflows with real email data.
- Gather information about extant email appraisal, transfer, processing, and access workflows.

In order to test the software in a variety of different computing environments, PREPARE opted to install and deliver the software using virtual machines (VM) constructed for this purpose. This method was chosen as it would create a uniform platform for the volunteers to be able to test the software with the same installations, configurations, and hardware prerequisites, with the hope that this would allow the results of the testing to be measured and compared beginning from the same baseline.

Ultimately, the project team developed two virtual machines for the testing phase:

- A Linux-based (Ubuntu 18.04 LTS) "primary" VM to serve as the default environment, with:
  - 8 GB RAM
  - 12 GB storage space (expandable to more, as was necessary)
  - ePADD, DArcMail, and libratom installed and configured according to the developers' specifications (wherever available)
  - Pre-configured file paths and command-line scripts or commands to operate the above tools (in short user guides)
  - Mozilla Thunderbird for transforming mail, if necessary
  - Links to developer documentation available for the above, as well as specific instructions on how to activate and operate tools within this VM
  - A location for email to be stored in the user's Home directory, to facilitate access
- A Windows 10 "secondary" VM to serve as an alternate environment for testing. Specifications were similar to the Linux VM but Windows-based versions of the tools were made available in this way.

Each of the VMs was re-issued (at the ⅓ and ⅔ mark of the testing period) as major updates to improve functionality, instructions, and organization of information; and to address feedback from the testing group. Minor updates and troubleshooting were available throughout the process.

Testers were given the option of using their own email datasets, or to use a PREPARE-provided email dataset for this purpose. The PREPARE-provided dataset featured over 14,000 emails originating from a single Utah state government user's account and made available to the project by the Utah State Archives. The dataset was provided as a nested hierarchy of folders roughly matching the user's account organization, with individual emails in EML format with attachments embedded in each folder and subfolder. As part of the design (and subsequent updates), the PREPARE Project Manager converted the EML files into a variety of different MBOX packages, some including all the emails in the dataset, others including only messages from each folder, with others creating single messages. Creating MBOX files was a necessary step in the process due to the input requirements of the software.

Notably, the sample dataset provided by the Utah State Archives was not distributed with the VM as part of the initial organization of the group, as both a space-saving measure and to better control the sharing of state datasets which was intended for this purpose only.

Volunteer Testing Group

A testing group was convened based originally on responses to the Needs Assessment Survey, including representatives from those states who specifically offered their assistance. Additionally, the PREPARE Project Manager contacted individual archivists who had participated in past grant projects (including the 2020 MoVE-IT project) to arrive at a group of 12 testers composed of archivists, digital archivists, and records managers from eight states and territories of varying size and digital preservation capability. Due to licensing restrictions, some of the proprietary software testing (specifically Emailchemy and CoolUtils'

PST Converter) was only able to be performed by the PREPARE Project Manager.

Testers were asked ahead of the testing phase whether they wanted to use the set of test email data provided by the Utah State Archives, or to use their own email datasets. Of the initial testing group, about half opted to bring their own data and the others opted to use the Utah dataset.[2] Over the course of the project, several more of the testers who had previously used their own email opted to test with the sample dataset as well.

Each member of the group was asked to contribute 2 hours per week during the three-month testing period, a total commitment of about 24 hours. In addition to testing the software for use in their workflows, each tester was asked to fill out a data collection survey for each of the tools tested. Each survey collected basic data about the dataset (for those not using the test dataset) and then quantitative and qualitative information about the tools and potential ways to use them. Measurements of the "usefulness" of each tool in various potential workflow steps (e.g. entity extraction, searching, organization, redaction, etc.) were taken to fill out the quantitative analyses, and narrative questions about how and when the tools could best be used, what features would be desired, and what limitations existed were gathered to measure the work qualitatively.

The PREPARE Project's goal is to gather information about *potential use of* tools in email processing at state and territorial archives, rather than the specific efficacy of each of these tools. Functionality was an important measure, but rather than "raw" functionality, testers were asked to evaluate possible use cases. Thus, it was important for the testers to have access to their own materials to provide as close to a "real" scenario as

VOLUNTEER TESTING GROUP MEMBERS

NAME	TITLE	STATE
Nick Connizzo	PREPARE Project Manager	CoSA
Carol Kussmann	Digital Records Analyst	(Univ of MN)
Rachel Smith	Collections Archivist	AL
Jessica Harden	State Government Archivist	MI
Jamie Patrick-Burns	Digital Archivist	NC
Erin Gallagher	Digital Description Archivist	NC
Gwen Amsbury	Records Management Analyst	OR
Amandeep Gogia	Records Management Analyst	OR
Roger Christman	Senior State Governor's Records Archivist	VA
Alan Arellano	State Records Archivist	VA
Krista Sorenson	Electronic Records Archivist	WI
Cindy Brown	Digital Archivist	WY

BASIC WORKFLOW

1. **Begin**
   - a. Direct the tool to the directory where the email dataset is located.
   - b. Select the email dataset to use (some or all).
2. **Appraisal/Survey**
   - a. Gather basic information regarding the size, scope, and content of the dataset.
   - b. Export/save this data and any other useful reports, evaluate their utility.
   - c. Search the dataset for sensitive information (or any other information of your choice).
3. **Processing**
   - a. Arrangement
     - i. Select and weed emails and attachments (as appropriate).
     - ii. Remove, segregate, and/or redact records with selected information.
     - iii. Organize records (as appropriate).
   - b. Description
     - i. Capture technical, administrative, structural, and preservation metadata (when possible).
     - ii. Add labels or metadata (as appropriate).
     - iii. Create other descriptive products.
4. **Outputs**
   - a. Report on actions taken, data removed or redacted, and other critical information.
   - b. "Export" completed collection in preferred formats.
   - c. Test usability of finalized records set (expected to be performed outside these tools)

2 However, some testers discovered that the limitations of the software (input formats had specific requirements) necessitated switching to use the test data instead of their own datasets. Specifically, email that has been preserved in TXT, HTML, or other file-formats not specifically designed for electronic messaging were not candidates for conversion to MBOX with available software. Email generated from older applications were especially troublesome.

possible.[3] Therefore, the simple software and environment-agnostic workflow (see page 3) was developed to guide the volunteer testers in working with the software.

The data collection survey gathered data about these key elements:

- **User information:** questions about the volunteer testing the tools (e.g., title, experience using processing tools, whether their institution had formal, established procedures for email processing)
- **Statistical information:** size and scope of the email datasets tested
- **Workflow:** measuring time spent on workflow steps, ideal processing steps, and any instances of data loss
- **Quantitative measures:** ratings of tool effectiveness for desired workflow steps including arrangement, audit trails, documentation, metadata, recovery, redaction, reporting scalability, search, and usability
- **Qualitative measures:** overall user experience, how software tools would fit in existing workflows, obstacles to implementation, and suggestions on how to improve functionality

At the close of the testing period, volunteers were asked to complete a survey for each tool tested. While each member of the group was assigned one tool to focus on, all were encouraged to test as many as they wished, and many were able to test multiple tools.

Challenges

The most significant challenge encountered was that most testers were not intimately familiar with the functioning of virtual machines. The PREPARE Project Manager provided instructional guides on how to perform some basic operations such as sharing a folder with the host machine or adding additional space to the VM to support larger batches of emails to process.

VMs were deployed in the Open Virtualization Format (OVA/OVF) to allow for interoperability regardless of the software used to run the VM. For testers, some used Oracle's VirtualBox and others used VMWare Workstation, and, during the testing phase, the project team discovered that the original VM did not function properly with the VMWare player due to format versioning issues (it was originally issued in OVF v2 and needed to be in v1 to function). Others had difficulty adding additional space and sharing folders. These issues led to the deployment of new VM versions to resolve them after extensive troubleshooting.

Installation and configuration of the tools was somewhat challenging as well. Each tool had some specific requirements in order to install them so that they would function. These tools were probably not intended for use in VMs, though the team was able to get all of them to function in each of the VMs deployed. It is still unclear, however, if some of the areas where the software had functionality issues were the result of improper installation and configuration, bugs in the software, or user error.

If creating another set of VMs for this purpose, they must be optimized with enough memory and processing power to be able to handle the operations of these software, especially when processing large collections.[4] A properly configured VM could serve as an open-source tools piloting testbed if designed with usability, repeatability, and stability being prioritized.

Results

Volunteers had strongly positive feelings about the tools' functionality. Each tester reported at least moderate success with one of the three tools, and most opted to try out all of the primary testing software.

All of the tools tested worked primarily with the MBOX format, but the set of test emails was preserved in EML format. While the volunteers were provided software to allow this conversion, the PREPARE Project Manager also made 3 different MBOX conversions from the master data set (Mozilla Thunderbird's Import/Export extension, CoolUtils' MBOX Converter, and Emailchemy's email converter). Most notably, each of the three produced slightly different results. This ended up causing problems as some of the MBOX files functioned well with ePADD yet would fail to load in DArcMail.

Migration of email from its native format or system (most likely Microsoft Exchange / Office 365 or Google Workspace) is likely to present the most significant challenge in the email preservation workflow.

It quickly became clear that there is a disconnect—most state governments use Microsoft Exchange for their email services, and thus generally output their email as PST archives or individual MSG files. Converting these to MBOX is not necessarily a simple process, and each conversion software appears to have its own quirks. Undoubtedly this is due to the closed-source nature of Microsoft's file formats and the difficulty capturing all the information contained in their proprietary formats when migrating to MBOX—some aspects which may be included in such archives just don't translate well to the MBOX format.

3 Among the testing group, there was substantial variance in the structure and format of existing email datasets for those who used their own email collections. One of the volunteers brought an email set to test in HTML format, which could not easily be converted back to EML or MBOX format.

4 There were several instances where collections over 2GB in size would hang or cause the VM to crash.

ePADD

Of the tools tested, ePADD provides the most fully-featured platform. It is deployed either as a standalone Windows application or individual modules in their own JAR containers.

Best for:

- A "one-stop shop" for appraising, processing, and providing access to high-value email collections
- Reviewing email and attachments, and searching for sensitive information
- Quick download, install, and run—5 minutes from download to actual work

Volunteers found ePADD fairly intuitive and easy-to-use. Ingest of emails is straightforward as is review of many elements found within. Testers primarily focused on the Appraisal and Processing modules, and generally favored the appraisal functionality. Perhaps most valuable was the wealth of data the software can extract and summarize regarding attachments, senders and recipients, labels, entities, and other details.

TESTER COMMENTS

"While the information provided by the tool, especially identifying the correspondents, entities, and other aggregating features was interesting and possibly useful, I was unsure what to do with all of it."

"What I was really interested in was identifying, segregating, and restricting access to any 'sensitive' data in the emails, and those steps got lost in all the other information. The interface to view the messages was good, but it was hard to know how to conduct any actions once I identified a message or content that I would want to restrict."

"It concerns me that I got slightly different results from two separate imports of the same MBOX files. The second import got slightly different counts of messages as well as sensitive information from the lexicon, which worries me. However, it bags the resulting files, which aligns with our use of bagging."

DArcMail

DArcMail is a medium-weight application designed with preservation of email as a core goal. It functions primarily by extracting the data from individual emails into a SQLite database from which additional functions can be performed.

Installation of DArcMail was the most difficult of the three primary tools, in part due to its prerequisites. Additionally, multiple versions of MBOX files created failed to load in DArcMail while they seemed to operate correctly when used with other software. It is possible that DArcMail has very specific format requirements (or perhaps DArcMail steers closely to a set of baseline requirements to which the conversion tools used in PREPARE didn't conform correctly).

Best for:

- Processing larger collections (not recommended for more than 100,000 messages)
- Transforming MBOX into XML (specifically EAXS[5]) or CSV

TESTER COMMENTS

"There were error messages every time I ran the tool, but they didn't seem to correspond with the messages not coming through. The collection that was almost 3GB failed to load."

"Good general metadata about the collection, but not a good way to export that info."

"Searches worked—when you knew what you wanted to search for. But if you don't know who/what to search for you don't have a starting place. To assist with that, I searched for blank spaces in the name field to get a list of names, then I could see more about who/what the messages were from and about."

"Searching for content was difficult in that you needed to know what you wanted to search for prior to searching. The process itself was easy."

"You seemed to have to know about the contents of the collection if you wanted to perform useful searches."

libratom

libratom is a command-line tool focused on extracting information from sets of email (especially PSTs) in order to facilitate the appraisal of email.

Volunteers found the tool to have the highest learning curve (lacking GUI), and requiring the most referral to the documentation. Users found the outputs interesting but had difficulty imagining how to incorporate those outputs into their email acquisition workflows. Users also lamented the inability to manipulate or interpret the tools within libratom, resulting in the need for additional software. Best for:

- Extraction of data from PST and MBOX
- Using natural language processing (NLP) to parse emails
- Conversion of PST/MBOX to EML

TESTER COMMENTS

"This is not a tool that you can install and go. It takes some time to figure out how to make it work and what you want to / can do with the results. The results within the tool itself are not super useful but can be used with other tools to analyze the results. It does run through a file quickly to show the number of messages within a file."

"This tool lacks the ability to really interpret the results within the tool itself."

"I liked that it worked with .pst files, but we also receive a lot of individual .msg files. It would be helpful if the tool could also work with those."

5 https://github.com/StateArchivesOfNorthCarolina/tomes-eaxs

Insights and Recommendations

As described in the 2021 *CoSA PREPARE: Needs Assessment Survey Analysis and Report,*

> *"best-practices and standards-oriented approach to enterprise email management could be broadly applicable to every jurisdiction. This model would focus on individual tasks and services to be performed to ensure emails are properly preserved at every point of their lifecycle, and that whoever needs to access them (the creator, IT, legal, HR, the public) has the proper, legal channels to do so."*[6]

Based on the survey responses from the volunteer group, the following products would make the greatest impact on the CoSA community for building email preservation capacity:

Standard Email Packaging Models

One of the most important data points to come out of the PREPARE Project testing is the wide variety of formats and structures already in place for extant collections of email currently being managed by state and territorial governments. It is much more challenging to build workflows that are effective and responsive when many different inputs must be considered. Standardizing formats and structures for email preservation (or even short-term management) can provide large leaps in efficiency in preserving email.

A technical approach to email governance will immensely aid archivists in appraising and processing emails. When archivists are unsure of the size, scope, format, and structure of incoming records, especially when there is little consistency across the enterprise of government, each incoming transfer must be treated as its own project with unique requirements. Archives will benefit greatly from the standardization of inputs as much as they will benefit from the standardization of outputs.

Software-Specific Workflows

CoSA, through its State Electronic Records Initiative, has been tracking the development of state and territorial archives' digital preservation maturity since 2012[7]. Results from the most recent Digital Preservation Capability Maturity Model self-assessment survey have demonstrated that maturity has been growing, especially in the areas over which archives have the most direct control: the growth of expertise and technology in their own institutions.[8] Digital archivists, including those in the testing group, have much more experience processing digital records than they did a decade ago.

Standardizing formats and structures for email preservation (or even short-term management) can provide large leaps in efficiency in preserving email.

Therefore, the efficacy of basic workflows, while still useful, is limited. Simple workflows, furthermore, run into the risks outlined in the previous point: they cannot necessarily be dropped into place in existing archival preservation frameworks. Digital archivists don't just need software to be available; they need direct, specific guidance on how best to incorporate these tools into their existing processes. Since email appraisal, processing, and discovery are fairly niche features, the few tools that do exist have fairly specific use-cases which may not be applicable to all governments. Therefore, specifically-designed and annotated workflows, especially those with clear inputs and outputs, could be useful products to rapidly improve email preservation.

Email Management Governance

As outlined in the *Needs Assessment Survey Analysis and Report*, there is a profound lack of email management governance in state and territorial government, leading to inconsistent approaches (at best) to the preservation of email. When empowered with the right tools, processes, and expertise, archivists should be able to more effectively communicate requirements for email preservation and incorporate them into existing email governance frameworks.

The management of email is almost certainly the responsibility of multiple agents throughout state and territorial governments: the email creator/recipient (the records producer) has the first responsibility, which then passes upward to the administrator of the email server(s) and associated infrastructure, and then onto the archives once records have met their disposition. Others may be involved as well: states with mature records management (RM) programs may have infrastructure that allows RM review/administration of email. Additionally, many jurisdictions have attorneys embedded in email-related review processes (especially redacting protected information).

6 Council of State Archivists. CoSA PREPARE: Needs Assessment Survey Analysis and Report. 2021. p. 8.

7 https://www.statearchivists.org/blogs/michelle-gallinger/2021/10/19/2022-dpcmm-history-of-the-dpcmm

8 Council of State Archivists. Self-Assessment Survey Report: 2022 Digital Preservation Capability Maturity Model. 2022.

A glossary of the responsibilities, requirements, or tasks associated with the proper management of email should be very valuable to all governments, for it can help to standardize the language and messages which must be disseminated, and more importantly absorbed and operationalized, throughout the enterprise of government. Outlining the tasks and potential roles related to email management can facilitate the assignment of responsibilities to the right personnel who can more effectively and efficiently accomplish such tasks. (For example, archivists *could* label and categorize individual emails but this task is almost certainly more effectively conducted by the records producer at the point of creation). By articulating what these tasks are, and what the archival requirements for the successful execution of such tasks might be, the archives can more effectively communicate to its designated communities what must be done to preserve email as it needs to be.

Digital archivists don't just need software to be available; they need direct, specific guidance on how best to incorporate these tools into their existing processes.

Conclusion

The PREPARE Project testing phase was extremely successful, with each of the testers getting the opportunity to test software for use in their email appraisal, acquisition, processing, and discovery workflows. CoSA is extremely grateful to the volunteers for their time and insights, to the state archives for granting us their staff time, to the State of Utah and the Utah State Archives for the use of their data set, and lastly the University of Illinois and the Mellon Foundation for their continued support of this project.

In Phase 3, the PREPARE Project moves to two objectives: utilize the information gathered and analyzed to inform the creation of templates, workflows, and other documentation, which will be concurrent with the crafting of a program to provide direct assistance to states and territories in their email preservation programs. Specifically, the PREPARE Project will seek to build archival capacity by designing custom workflows that leverage tools and techniques tested here, to develop governance for the high-level management and preservation of email, and to build and contribute to communities of practice by contributing to the development of open-source software and connecting members of the state and territory archives community working on email preservation.